RESEARCH CENTRE

Inria Saclay Center

2022 ACTIVITY REPORT

Project-Team SODA

Computational and mathematical methods to understand health and society with data

DOMAIN

Digital Health, Biology and Earth

THEME

Computational Neuroscience and Medicine

Inria

Contents

Project-Team SODA

Creation of the Project-Team: 2022 March 01

Keywords

Computer sciences and digital sciences

A3.3. – Data and knowledge analysis

A3.4. – Machine learning and statistics

A9.1. – Knowledge

A9.2. – Machine learning

Other research topics and application domains

B2.3. – Epidemiology

B9.1. – Education

B9.5.6. – Data science

B9.6.1. – Psychology

B9.6.3. – Economy, Finance

1 Team members, visitors, external collaborators

Research Scientists

- Gael Varoquaux [Team leader, INRIA, Senior Researcher, HDR]
- Judith Abecassis [INRIA, Researcher]
- Marine Le Morvan [INRIA, Researcher]
- Jill Jenn Vie [INRIA, Researcher]

Post-Doctoral Fellows

- Riccardo Cappuzzo [INRIA, from Jul 2022]
- Myung Kim [INRIA, from Nov 2022]

PhD Students

- Samuel Brasil De Albuquerque [INSERM]
- Lihu Chen [Telecom ParisTech]
- Bénédicte Colnet [INRIA]
- Alexis Cvetkov-Iliev [INRIA]
- Matthieu Doutreligne [HAS]
- Leo Grinsztajn [INRIA]
- Alexandre Perez [INRIA]

Technical Staff

- David Arturo Amor Quiroz [INRIA, Engineer]
- Franck Charras [INRIA, Engineer, from Jul 2022]
- Jeremie Du Boisberranger [INRIA, Engineer]
- François Goupil [INRIA, Engineer]
- Olivier Grisel [INRIA, Engineer]
- Julien Jerphanion [INRIA, Engineer]
- Guillaume Lemaitre [INRIA, Engineer]
- Vincent Maladiere [INRIA, Engineer, from Aug 2022]
- Tomas Rigaux [INRIA, Engineer]
- Jovan Stojanovic [INRIA, Engineer, from Mar 2022]

Interns and Apprentices

- Lilian Boulard [INRIA, Apprentice]
- Anand-Arnaud Pajaniradjane [CENTRALESUPELEC, Intern, from Nov 2022]

Visiting Scientist

- Ko Takeuchi [Kyoto University, from Oct 2022]

2 Overall objectives

2.1 Context

2.1.1 Application context: richer data in health and social sciences

Opportunistic data accumulations, often observational, bare great promises for social and health sciences. But the data are too big and complex for standard statistical methodologies in these sciences.

Health databases Increasingly rich health data is accumulated during routine clinical practice as well as for research. Its large coverage brings new promises for public health and personalized medicine, but it does not fit easily in standard biostatistical practice because it is not acquired and formatted for a specific medical question.

Social, educational, and behavioral sciences Better data sheds new light on human behavior and psychology, for instance with on-line learning platforms. Machine learning can be used both as a model for human intelligence and as a tool to leverage these data, for instance improving education.

2.1.2 Related data-science challenges

Data management: preparing dirty data for analytics Assembling, curating, and transforming data for data analysis is very labor intensive. These data-preparation steps are often considered the number one bottleneck to data-science. They mostly rely on data-management techniques. A typical problem is to establishing correspondences between entries that denote the same entities but appear in different forms (entity linking, including deduplication and record linkage). Another time-consuming process is to join and aggregate data across multiple tables with repetitions at different levels (as with panel data in econometrics and epidemiology) to form a unique set of "features" to describe each individual. This process is related to database denormalization.

Progress in machine learning increasingly helps automating data preparation and processing data with less curation.

Data science with statistical machine learning Machine learning can be a tool to answer complex domain questions by providing non-parametric estimators. Yet, it still requires much work, eg to go beyond point estimators, to derive non-parametric procedures that account for a variety of bias (censoring, sampling biases, non-causal associations), or to provide theoretical and practical tools to assess validity of estimates and conclusion in weakly-parametric settings.

3 Research program

3.1 Representation learning for relational data

Soda develops develop deep-learning methodology for relational databases, from tabular datasets to full relational databases. The stakes are *i)* to build machine-learning models that apply readily to the raw data so as to minimize manual cleaning, data formatting and integration, and *ii)* to extract reusable representations that reduce sample complexity on new databases by transforming the data in well-distributed vectors.

3.2 Mathematical aspects of statistical learning for data science

While complex models used in machine learning can be used as non-parametric estimators for a variety of statistical task, the statistical procedures and validity criterion need to be reinvented. Soda contributes statistical tools and results for a variety of problems important to data science in health and social science (epidemiology, econometrics, education). Statistical topics of interest comprise:

- Missing values
- Causal inference
- Model validation
- Uncertainty quantification

3.3 Machine learning for health and social sciences

Soda targets applications in health and social sciences, as these can markedly benefit from advanced processing of richer datasets, can have a large societal impact, but fall out of mainstream machine-learning research, which focus on processing natural images, language, and voice. Rather, data surveying humans needs another focus: it is most of the time tabular, sparse, with a time dimension, and missing values. In term of application fields, we focus on the social sciences that rely on quantitative predictions or analysis across individuals, such as policy evaluation. Indeed, the same formal problems, addressed in the two research axes above, arise across various social sciences: **epidemiology, education research, and economics**. The challenge is to develop efficient and trustworthy machine learning methodology for these high-stakes applications.

3.4 Turn-key machine-learning tools for socio-economic impact

Societal and economical impact of machine learning requires easy-to-use practical tools that can be leveraged in non-specialized organizations such as hospitals or policy-making institutions.

Soda incorporates the core team working at Inria on **scikit-learn**, one of the most popular machine-learning tool world-wide. One of the missions of soda is to improve scikit-learn and its documentation, transfering the understanding of machine learning and data science accumulated by the various research efforts.

Soda works on other important software tools to foster growth and health of the Python data ecosystem in which scikit-learn is embedded.

4 Application domains

4.1 Precision medicine, public health, and epidemiology

Data management is the focus of the field of medical informatics as it is notably challenging in healthcare settings, due to the multiplicity of sources and the richness of the data that encompasses many modalities. We apply the our machine techniques for statistical analysis, including causal inference, in medicine to facilitate clinical research. The central questions are that of personalized medicine –prediction at the individual level, for diagnosis, prognosis, or drug recommendation– and of public health –evaluation of treatments and policy, estimation of risk factors. The data on which we work are patient history and claims databases: mid-dimensional data with longitudinal coverage (as opposed to “omics” or imaging data, which is high dimensional and much less frequently available in clinical settings).

We collaborate actively with AP-HP and Haute Autorité de Santé. APHP provides access to its very rich and complex data mart, with thousands of tables following millions of individuals, both a challenge and an opportunity, and we work with various medical specialists (neurology, diabetology, public health) on specific clinical questions related to prognostic, treatment evaluation, and risk factors. Haute Authorité de Santé collaborates with us by to answer public-health questions. These typically require causal inference. Finally, we are in close discussions with Institut Curie and HDH (health data hub), to start

collaborations on respectively evaluation of oncology treatments using electronic health records and automatic featurization in health databases.

4.2 Educational data mining

In educational data mining, we are interested in developing mathematical methods of learning to personalize education through adaptive assessment (developing algorithms that select questions for measuring efficiently the latent knowledge of examinees or for optimizing learning), recommending learning resources, generating exercises automatically. It is a challenging problem as it is hard to quantify learning, unlike in traditional reinforcement learning scenarios, and it is hard to measure the effect of courses on learning. This is why it is traditionally modeled as a partially-observable Markov decision process (POMDP). We are interested in modeling the evolution of uncertainty over the latent knowledge of examinees over time, for example using Bayesian approches.

Soda has deep historical ties with the national platform pix.fr for certifying the digital competencies of all French citizens. Jill-Jênn Vie is one of the original core developers and organized a workshop with them in 2020. Another source of data on education, besides public datasets for example from Duolingo, comes from our collaboration with NeuroSpin, where Stanislas Dehaene directs the scientific council of national education (Conseil Scientifique de l'Éducation nationale) giving access to statistics from DEPP (direction de l'évaluation, de la prospective et de la performance) and evaluation data nationwide. We are in ongoing discussions with ONISEP (Office national d'information sur les enseignements et les professions) to get orientation data about French high schoolers and documentation/interviews about jobs.

4.3 Data management

Data preparation for analytics is intrinsically related to data management. For instance, linked open data provides consistent views on data across silos, but integrating these data into a statistical model to answer a given question still requires a lot of user efforts. Database operation increasingly relies on machine learning. While Soda is in no way expert in database research, the analytic tools that we build for relational data are increasingly used for data management.

4.4 Broader data science

The tools, practical and theoretical, that we develop are central to many applications of data science. For instance, we often discuss with banks and insurances, which use machine learning but face statistical problems that we tackle: censoring or other sampling biases, forecasting, uncertainty quantification. Marketing and business intelligence also face the same exact problems. Even more generally, data preparation from relational databases is a challenge is most data-science applications. We interact with data scientists in a broad set of applications via the user base of the software tools that we develop (*eg* scikit-learn) and the various courses and lectures that we give around these tools to industry audiences.

4.5 Behavioral sciences

A methodological challenge in health and educational sciences common to behavioral science is that the quantities of interest are difficult to measure, eg intelligence or progress of a student. Supervised machine learning can infer proxies from indirect signs, such as psychological traits from brain imaging, diagnosis from clinical traces, or socio-economical status from demographics. This notion of proxies is central in policy evaluation, serving as indirect signals in causal inference, to provide secondary outcomes for treatment effect estimation or to control confounds not directly observed.

5 Social and environmental responsibility

5.1 Footprint of research activities

The main footprint of Soda's activity is the carbon footprint of our travels (surpassing our compute cost, as we seldom run very intensive computation). For this reason, we try to be careful with our long-distance travel and try to take the plane as little as possible. Not flying at all is not possible, as it would cut us off from the world-wide research community sometimes mediated by crucial conferences in North America. However, we favor online seminars, or on-premise talks accessible by train.

5.2 Impact of research results

While data science can improve health and education, working with personal data or providing decision tools that affect individuals comes with responsibilities.

First, overly optimistic claims, improper evaluations, or faulty statistical analysis can lead to premature usage of personal data. As collecting and handling personal data comes with privacy risks, a sober analysis of cost-benefit trade-offs is good practice. We strive to develop methodology for good evaluation [23, 24] and raise awareness of pitfalls [3].

Second, Soda does not put any tools in production: none of the works of soda directly leads to automated decisions. Consequently none of our work has directly impacted individuals.

Finally, Soda works on pseudonymized data, and we leave the –pseudonymized– electronic health data on servers inside the protected environment of the hospital where they have been acquired and are used. Going further, Soda runs research on privacy-preserving synthetic data generation, to provide open datasets for research and development without privacy concerns [22].

6 Highlights of the year

6.1 General News

Soda was created in March 2022.

Judith Abecassis was recruited as a PI in September 2022.

6.2 Awards

Gaël Varoquaux highly-cited clarivate researcher Gaël Varoquaux is on the list of the highly-cited clarivate researchers for year 2022.

Prix de la science ouverte du logiciel libre `scikit-learn` received an award from the French ministry of higher education on open-source scientific software.

7 New software and platforms

7.1 New software

7.1.1 Scikit-learn

Keywords: Clustering, Classification, Regression, Machine learning

Scientific Description: Scikit-learn is a Python module integrating classic machine learning algorithms in the tightly-knit scientific Python world. It aims to provide simple and efficient solutions to learning problems, accessible to everybody and reusable in various contexts: machine-learning as a versatile tool for science and engineering.

Functional Description: Scikit-learn can be used as a middleware for prediction tasks. For example, many web startups adapt Scikitlearn to predict buying behavior of users, provide product recommendations, detect trends or abusive behavior (fraud, spam). Scikit-learn is used to extract the

structure of complex data (text, images) and classify such data with techniques relevant to the state of the art.

Easy to use, efficient and accessible to non datascience experts, Scikit-learn is an increasingly popular machine learning library in Python. In a data exploration step, the user can enter a few lines on an interactive (but non-graphical) interface and immediately sees the results of his request. Scikitlearn is a prediction engine . Scikit-learn is developed in open source, and available under the BSD license.

URL: http://scikit-learn.org

Publications: hal-00650905, hal-00856511, hal-01093971

Contact: Olivier Grisel

Participants: Alexandre Gramfort, Bertrand Thirion, Gael Varoquaux, Loic Esteve, Olivier Grisel, Guillaume Lemaitre, Jeremie Du Boisberranger, Julien Jerphanion

Partners: Boston Consulting Group - BCG, Microsoft, Axa, BNP Parisbas Cardif, Fujitsu, Dataiku, Assistance Publique - Hôpitaux de Paris, Nvidia

7.1.2 joblib

Keywords: Parallel computing, Cache

Functional Description: Facilitate parallel computing and caching in Python.

URL: https://joblib.readthedocs.io/en/latest/

Contact: Gael Varoquaux

7.1.3 dirty-cat

Keyword: Machine learning

Functional Description: Vectorizes tables with badly formated entries to enables statistical learning

URL: https://dirty-cat.github.io/stable/

Contact: Gael Varoquaux

8 New results

8.1 Representation learning for relational data

Participants: Gael Varoquaux.

Aggregating many tables into features [9] For many machine-learning tasks, augmenting the data table at hand with features built from external sources is key to improving performance. For instance, estimating housing prices benefits from background information on the location, such as the population density or the average income.

Most often, a major bottleneck is to assemble this information across many tables, requiring time and expertise from the data scientist. We propose vectorial representations of entities (e.g. cities) that capture the corresponding information and thus can replace human-crafted features [9]. We represent the relational data on the entities as a graph and adapt graph-embedding methods to create feature vectors for each entity. We show that two technical ingredients are crucial: modeling well the different relationships between entities, and capturing numerical attributes. We adapt knowledge graph embedding methods

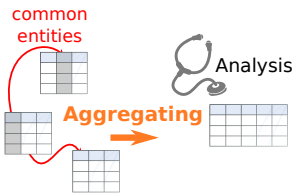

Figure 1: Often, data must be assembled across multiple tables into a single table for analysis. Challenges arise due to one-to-many relations, irregularity of the information, and the number of tables that may be involved.

that were primarily designed for graph completion. Yet, they model only discrete entities, while creating good feature vectors from relational data also requires capturing numerical attributes. For this, we introduce KEN: Knowledge Embedding with Numbers. We thoroughly evaluate approaches to enrich features with background information on 7 prediction tasks. We show that a good embedding model coupled with KEN can perform better than manually handcrafted features, while requiring much less human effort. It is also competitive with combinatorial feature engineering methods, but much more scalable. Our approach can be applied to huge databases, for instance on general knowledge graphs as in YAGO, creating general-purpose feature vectors reusable in various downstream tasks (Figure 2).

Imputing out-of-vocabulary embeddings with LOVE [18] Modern natural language processing systems represent inputs with word embeddings. Likewise, analytics on relational data can be built with entity embeddings, as above. However, these approach are brittle when faced with Out-of-Vocabulary (OOV) words or entities. To address this issue, we follow the principle of mimick-like models to generate vectors for unseen words, by learning the behavior of pre-trained embeddings using only the surface form of words [18]. We present a simple contrastive learning framework, LOVE (Learning Out of Vocabulary Embeddings), which extends the word representation of an existing pre-trained language model (such as BERT), and makes it robust to OOV with few additional parameters. Extensive evaluations demonstrate that our lightweight model achieves similar or even better performances than prior competitors, both on original datasets and on corrupted variants. Moreover, it can be used in a plug-and-play fashion with FastText and BERT, where it significantly improves their robustness.

Tabular machine learning [19] While deep learning has enabled tremendous progress on text and image datasets, its superiority on tabular data is not clear. We contribute extensive benchmarks [19] of standard and novel deep learning methods as well as tree-based models such as XGBoost and Random Forests, across a large number of datasets and hyperparameter combinations. We define a standard set of 45 datasets from varied domains with clear characteristics of tabular data and a benchmarking methodology account- ing for both fitting models and finding good hyperparameters. Results show that tree-based models remain state-of-the-art on medium-sized data (~ 10 K samples) even without accounting for their superior speed. To understand this gap, we conduct an empirical investigation into the differing inductive biases of tree-based models and neural networks. This leads to a series of challenges which should guide researchers aiming to build tabular-specific neural network: *1)* be robust to uninformative features, *2)* preserve the orientation of the data, and *3)* be able to easily learn irregular functions. To stimulate research on tabular architectures, we contribute a standard benchmark and raw data for baselines: every point of a 20 000 compute hours hyperparameter search for each learner. The conclusion that tree-based learners outperform deep learning on tabular data is interesting from a resource standpoint: these are indeed much more frugal in resources.

8.2 Mathematical aspects of statistical learning for data science

Participants: Marine Le Morvan, Gael Varoquaux.

Validating probabilistic classifiers: beyond calibration [23] Ensuring that a classifier gives reliable confidence scores is essential for informed decision-making. For instance, before using a clinical prognostic model, we want to establish that for a given individual is attributes probabilities of different clinical

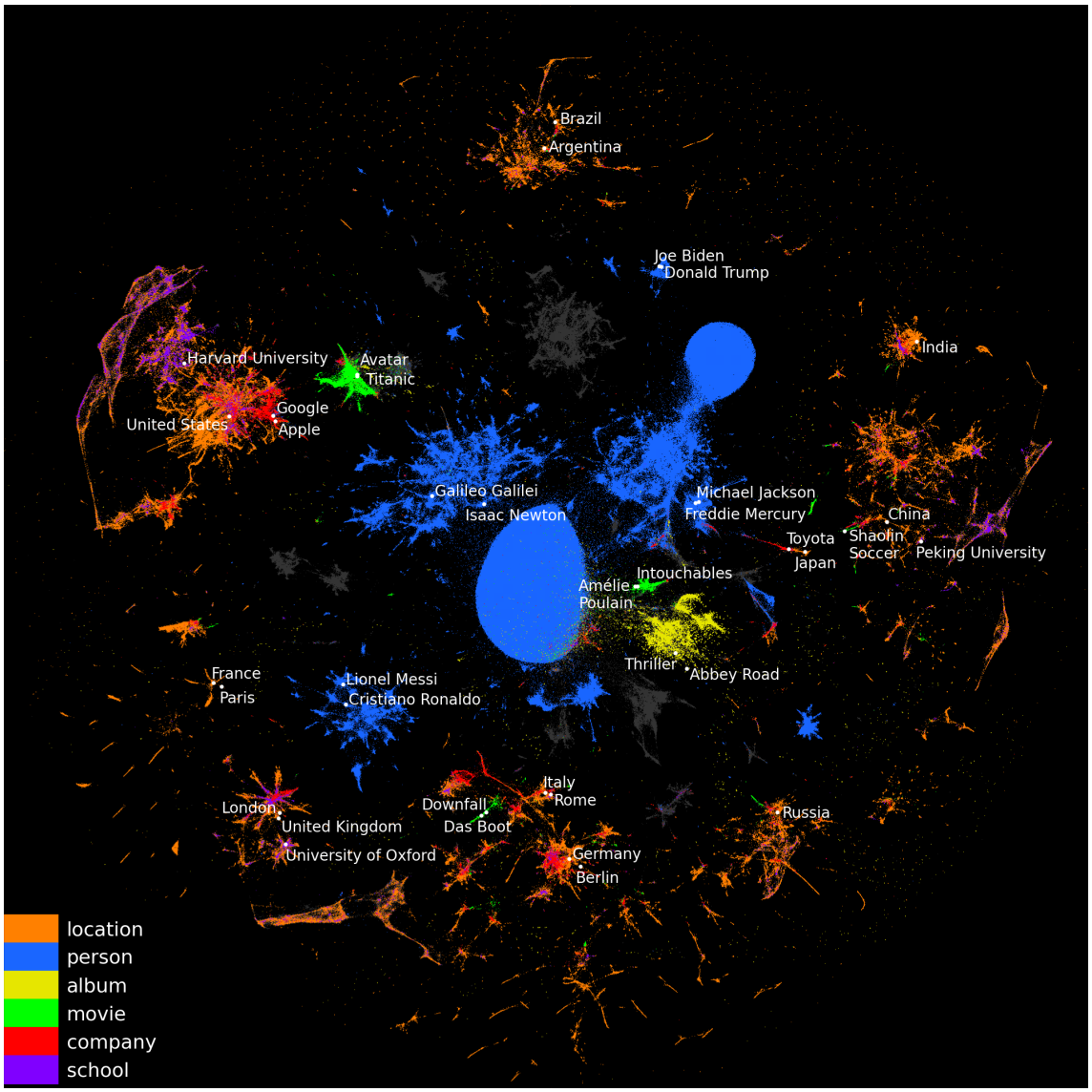

Figure 2: 2D-representation (using UMAP) of the entity embeddings of YAGO (wikipedia). The vectors are downloadable from soda-inria.github.io/ken_embeddings to readily augment data-science projects.

outcomes that can be indeed trusted. To this end, recent work has focused on miscalibration, *i.e.*, the over or under confidence of model scores. Yet calibration is not enough: even a perfectly calibrated classifier with the best possible accuracy can have confidence scores that are far from the true posterior probabilities, if it is over-confident for some samples and under-confident for others. This is captured by the grouping loss, created by samples with the same confidence scores but different true posterior probabilities. Proper scoring rule theory shows that given the calibration loss, the missing piece to characterize individual errors is the grouping loss. While there are many estimators of the calibration loss, none exists for the grouping loss in standard settings. We propose an estimator to approximate the grouping loss [23]. We show that modern neural network architectures in vision and NLP exhibit grouping loss, notably in distribution shifts settings, which highlights the importance of pre-production validation.

Causal inference: handling missing covariates when generalizing to new populations [7] Randomized Controlled Trials (RCTs) are often considered as the gold standard to conclude on the causal effect of

a given intervention on an outcome, but they may lack of external validity when the population eligible to the RCT is substantially different from the target population: due to sampling biases they measure on the study population an effect different than that of the target population. Having at hand a sample of the target population of interest allows to generalize the causal effect. Identifying this target population treatment effect needs covariates in both sets to capture all treatment effect modifiers that are shifted between the two sets. However such covariates are often not available in both sets. Standard estimators then use either weighting (IPSW), outcome modeling (G-formula), or combine the two in doubly robust approaches (AIPSW). In this work, after completing existing proofs on the complete case consistency of those three estimators, we computed the expected bias induced by a missing covariate, assuming a Gaussian distribution and a semi-parametric linear model. This enables sensitivity analysis for each missing covariate pattern, giving the sign of the expected bias. We also showed that there is no gain in imputing a partially-unobserved covariate. Finally we studied the replacement of a missing covariate by a proxy. We illustrated all these results on simulations, as well as semi-synthetic benchmarks using data from the Tennessee Student/Teacher Achievement Ratio (STAR), and with a real-world example from critical care medicine.

8.3 Machine learning for health and social sciences

Participants: Gael Varoquaux, Jill-Jênn Vie.

Challenges to clinical impact of AI in medical imaging [17] Research in computer analysis of medical images bears many promises to improve patients' health. However, a number of systematic challenges are slowing down the progress of the field, from limitations of the data, such as biases, to research incentives, such as optimizing for publication. We reviewed roadblocks to developing and assessing methods. Building our analysis on evidence from the literature and data challenges, we showed that at every step, potential biases can creep in [17]. First, larger datasets do not bring increased prediction accuracy and may suffer from biases. Second, evaluations often miss the target, with evaluation error larger than algorithmic improvements, improper evaluation procedures and leakage, metrics that do not reflect the application, incorrectly chosen baselines, and improper statistics. Finally, we show how publishing too often leads to distorted incentives. On a positive note, we also discuss on-going efforts to counteract these problems and provide recommendations on how to further address these problems in the future.

Privacy-preserving synthetic educational data generation [22] Institutions collect massive learning traces but they may not disclose it for privacy issues. Synthetic data generation opens new opportunities for research in education. We presented a generative model for educational data that can preserve the privacy of participants, and an evaluation framework for comparing synthetic data generators. We show how naive pseudonymization can lead to re-identification threats and suggest techniques to guarantee privacy. We evaluate our method on existing massive educational open datasets.

8.4 Turn-key machine-learning tools for socio-economic impact

Participants: Olivier Grisel, Guillaume Lemaitre, Gael Varoquaux.

New releases of scikit-learn Scikit-learn is always improving, adding features for better and easier machine learning in Python. We list below a few highlights that are certainly not exhaustive but illustrate the continuous progress made.

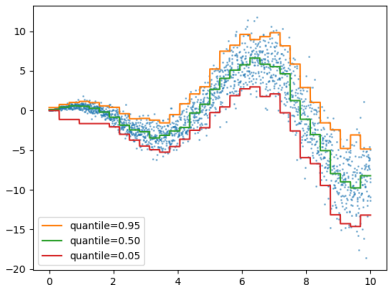

Figure 3: Quantile loss in `HistGradientBoostingRegressor`

Release 1.1 (may 2022)

- Quantile loss in `HistGradientBoostingRegressor`, to estimate conditional quantile.
- Grouping infrequent categories in `OneHotEncoder`
- `MiniBatchNMF`: an online version of NMF (non-negative matrix factorization, much more scalable.
- `BisectingKMeans`: divide and cluster for more regular clusters than normal `KMeans`.
- Improved efficiency of many estimators. The efficiency of estimators relying on the computation of pairwise distances (essentially estimators related to clustering, manifold learning and neighbors search algorithms) was greatly improved for float64 dense input. Efficiency improvement especially were a reduced memory footprint and a much better scalability on multi-core machines.
- Output feature names available in all transformers.

Release 1.2 (Dec 2022)

- Pandas output: all transformers (and thus all intermediate steps) can represent data as pandas dataframe, thus attaching relevant names to the various features and providing a data structure that is familiar to many users.
- Interaction constraints in Histogram-based Gradient Boosting Trees.
- New and enhanced visualization: `PredictionErrorDisplay` provides a way to analyze regression models in a qualitative manner; `LearningCurveDisplay` can more easily plots learning curves
- Faster parser in data downloader (from openml).
- Experimental GPU support, using the generalized array API in `LinearDiscriminantAnalysis` which opens the door to using cuda via CuPy.
- Improved efficiency of many estimators. The efficiency of many estimators relying on the computation of pairwise distances (essentially estimators related to clustering, manifold learning and neighbors search algorithms) was further improved for all combinations of dense and sparse inputs on float32 and float64 datasets, except the sparse-dense and dense-sparse combinations for the Euclidean and Squared Euclidean Distance metrics.

dirty-cat Dirty-cat is a much younger package that strives to facilitate statistical learning on relational data with poorly-normalized entries.

Release 0.3 (Sep 2022)

- The `SuperVectorizer` (to vectorize a table) is now suitable for automatic usage:
  - automatic casting of types in transform,
  - avoid dimensionality explosion when a feature has two unique values, by using a `OneHotEncoder` that drops one of the two vectors.
  - transform can now return features, without modification.
- New encoder: DatetimeEncoder can transform a datetime column into several numerical columns (year, month, day, hour, minute, second, ...). It is now the default transformer used in the `SuperVectorizer` for datetime columns.

joblib joblib is a very simple computation engine in Python that is used by many packages, including scikit-learn for parallel computing.

Release 1.2 (Sep 2022)

- Fix a security issue (potential code ingestion).
- Make joblib work on exotic architectures, such as Pyodide for computation in the browser using web assembly.
- Make sure that persistence respects memory alignment.

9 Bilateral contracts and grants with industry

Participants: Gael Varoquaux.

9.1 Bilateral contracts with industry

scikit-learn consortium `scikit-learn` is development is funded via a consortium with industry partners hosted by the Inria foundation. The consortium currently comprises hugginface, dataiku, BNP-Parisbas-cardiff, AXA, nvidia, BCG gamma, microsoft, and fujitsu, It has an operating budget of around 250 k€ yearly, and employs 5 engineers. Priorities are set by a bi-annual technical committee where industry partners sit together with community members in a joint discussion. Gaël Varoquaux is in charge at Soda.

Intel donation Intel has partnered with Soda in the context of its "oneAPI centers of excellence". The donation, an amount of 100 k€ yearly for two years, aims to establish a high-performance computational back-end to speed up scikit-learn on GPUs using `SYCL`. The backend is implemented as a plugin external to the scikit-learn codebase that can extend it. Gaël Varoquaux is in charge at Soda.

Collaboration with Haute Autorité de Santé We have a 3-year long collaboration with Haute Autorité de Santé (HAS) on using health data mart for policy evaluation. The collaboration is mediated by Matthieux Doutreligne –part time HAS, part time Soda– and comes with a budget of 50 k€. Gaël Varoquaux is in charge at Soda.

9.2 Bilateral Grants with Industry

Cython+ Cython+ is a funding by BPI-France and région Ile de France to develop parallel computing in Python. It united the software editors Nexedi, Albilian, and Mines-Telecom (IMT) and Inria. Inria was funded 100 k€ to develop parallel computing in `scikit-learn`. The funding ended end of August 2022. Gaël Varoquaux is in charge at Soda.

Plan de relance with Dataiku – Soda has a 24 months post-doc funded by "plan de relance" jointly with Dataiku on using embeddings for database analytics. The post-doc started beginning of November 2022. Gaël Varoquaux is in charge at Soda.

10 Partnerships and cooperations

Participants: Judith Abecassis, Marine Le Morvan, Gael Varoquaux, Jill-Jênn Vie.

10.1 International initiatives

10.1.1 Associate Teams in the framework of an Inria International Lab or in the framework of an Inria International Program

OPALE

Title: Optimal policy active learning for education

Duration: 2021 -> 2023

Coordinators: Jill-Jênn Vie and Koh Takeuchi (takeuchi@i.kyoto-u.ac.jp)

Partners:

- Kyoto University Kyoto (Japon)

Inria contact: Jill-Jênn Vie

Summary: Our research project is to explore reinforcement learning and causal inference to learn policies for collecting student data, so that we can understand how the students learn, and which lessons/exercises in a course have a strong impact on learning for which students. This has benefits both for the students that can reflect on their knowledge, and the teachers that can receive feedback and recommendations to improve their teaching: for example, if the teacher hesitates between several learning resources to master a topic, our algorithm would be able to quantify which lesson is most suitable for a given student. We plan to implement our algorithms in actual platforms to try the learned policies on real students, with public and private partnerships (for example Pix, a project initially based at the French Ministry of Education, which is now the public standardized test for certifying digital competencies in France).

10.2 International research visitors

10.2.1 Visits of international scientists

Other international visits to the team

Koh Takeuchi

Status assistant professor

Institution of origin: Kyoto University

Country: Japan

Dates: October 21–25

Context of the visit: OPALE associated team

Mobility program/type of mobility: research stay

10.2.2 Visits to international teams

Research stays abroad

Bénédicte Colnet

Visited institution: University of Berkeley & Stanford University

Country: USA

Dates: March – June

Context of the visit: Collaboration on causal inference

Mobility program/type of mobility: research stay

Lilian Boulard

Visited institution: TU Eindhoven

Country: The Netherlands

Dates: August

Context of the visit: Collaboration on encoding dirty categories

Mobility program/type of mobility: research stay

Jill-Jênn Vie

Visited institution: Kyoto University

Country: Japan

Dates: November–December

Context of the visit: OPALE associated team

Mobility program/type of mobility: research stay

10.3 European initiatives

10.3.1 Horizon Europe

Intercept-T2D project Soda is partner of the Intercept project, funded via the Horizon Europe Framework Programme (HORIZON), on *early interception of inflammatory-mediated Type 2 Diabetes*.

The project investigates whether an inflammatory-mediated profile contributes to the onset of Type 2 Diabetes (T2D) complications, thus enabling the identification of patients most at risk of complications and the design of personalized prevention measures. T2D is a heterogeneous disease, which is an obstacle to the delivery of an optimal tailored treatment. Consequently, patients' individual trajectories of progressive hyperglycemia and risk of chronic complications are so far difficult to predict. In this context, onset of diabetic complications represents the most important transitional phase of T2D development toward premature disability and mortality. Chronic systemic inflammation has been suggested to be a major contributor to the onset and progression of T2D complications. INTERCEPT-T2D will consider all diagnosis inflammatory parameters that are of importance for the transition to T2D-related complications. The combination of state-of-the-art genomics and cell-biology technologies with targeted clinical interventions should lead to potent patients' stratification, identification and prognosis of a novel class or subclass of patients characterized by "Inflammatory-mediated T2D".

The project will combine longitudinal human European cohorts of patients with an extensive health data warehouse, to establish the inflammatory trajectory of citizens with T2D from diagnosis to the development of complications. Soda focuses on exploiting the health data warehouse (namely APHP's EDS – entrepot de données de santée).

The project is coordinated by INSERM, and partners are: INRIA, the DDFG (Deutsche Diabetes Forschungs Gesellschaft), the Universita degli studi di Verona, the Karolinska institutet, the Technische Universitaet Dresden, APHP, CHU Liege, Federation Francaise des Diabetiques. Gaël Varoquaux is in charge at Soda.

10.4 National initiatives

IPL HPC / big-data Soda is part of the "Inria Project Lab" on HPC / big-data, a nation-wide collaboration across all Inria centers bridging research groups expert in high-performance computing with research on IA and data science. The project has been running for 5 years and is ending end of 2022. Gaël Varoquaux is in charge at Soda.

DirtyData project DirtyData is a research project on developing machine learning for non-curated datasets, funded by ANR since 2018 (ended in march 2022), with partners Telecom Paristech (Fabian Suchanek), CNRS (Balazs Kegl), Inria (Soda) and the company Data-publica. The main research axes have been machine learning with missing-values and machine learning on non-normalized relational data. Gaël Varoquaux is principal investigator of DirtyData.

10.5 Regional initiatives

MissingBigData MissingBigData is a research project on machine learning with missing values, funded by the DataIA institute since 2019 (ended in august 2022). The partners are Julie Josse (École polytechnique and Inria Montpellier) and Soda (Gaël Varoquaux). The research has focused on understanding what aspects of a learning architecture are important to handle missing values at train and at test time.

11 Dissemination

Participants: Judith Abecassis, Marine Le Morvan, Gael Varoquaux, Jill-Jênn Vie.

11.1 Promoting scientific activities

11.1.1 Scientific events: organisation

Member of the organizing committees

François Goupil Open-source experience, track leader

11.1.2 Scientific events: selection

Member of the conference program committees

Gaël Varoquaux NeurIPS: area chair & workshop reviewer

Reviewer

Gaël Varoquaux ICML, ICLR, AIstats, AAAI, IJCAI (member of the Program Committee Board – PCB)

Marine Le Morvan ICML, Neurips, ICLR.

11.1.3 Journal

Member of the editorial boards

Gaël Varoquaux elife, machine learning journal

Jill-Jênn Vie Journal of Educational Data Mining (JEDM)

Reviewer - reviewing activities

Gaël Varoquaux JMLR (journal of machine learning research), PLOS Digital Health

Marine Le Morvan Journal of Machine Learning Research (JMLR), Machine Learning journal (MACH)

Jill-Jênn Vie Journal of Educational Data Mining (JEDM)

11.1.4 Invited talks

Gaël Varoquaux

- AI4AD workshop@IEEE WCCI 2022, Artificial Intelligence for Alzheimer's Disease, online, July 22th, *Measuring mental health with machine learning and brain imaging*
- Seminar at Danish Technical University computer science department, Copenhaggen, Aug 18th, *Reinventing data science for statistical evidence from a messy world*
- Séminaire des élèves, Mastère mathématiques de l'IA, Orsay, Sept 22nd, *Supervised learning with missing values*
- Danish Data Science Academy, online, Nov 7th, *Scikit-learn: democratizing machine learning*
- DGKN (Deutsche Gesellschaft für Klinische Neurophysiologie – congrès annuel de la neurophysiologie clinique allemande), Wurzburg, March 11th, *Measuring mental health with machine learning and brain imaging*
- Euroscipy, Basel, Thur 1st, *Machine-learning with missing values*
- Journée de la Science ouverte, Grenoble, Dec 13th, *Logiciel libre et science ouverte dans la cité*
- Inria Project Lab Big-data / HPC, Paris, May 30th, *scikit-learn performance work*
- Working "Learning to discover", Orsay, Apr 27th, *Scientific inference with imperfect theories, Examples with machine learning and neurosciences*
- Leuven.AI keynote, Leuven, Jun 2nd, *DirtyData: statistical learning on non-curated databases*
- Journée reproductibilité labex primes, Lyon, Dec 8th, *Without meaningful benchmarks, machine learning research is vain*
- Seminar CSAIL (Computer Science AI Lab), MIT, Boston, May 25th, *DirtyData: statistical learning on non-curated databases*
- Keynote at Promenta conference, Oslo (remote participation), Sept 28th, *Measuring mental health with machine learning and brain imaging*
- Seminar Max Planck Institute fur Informatiks, Saarbruken, March 9th, *Embeddings of databases for analytics robust to alignment errors*
- Keynote DCASE, Nancy, Nov 3rd, *Model evaluation, a machine-learning bottleneck*
- OHBM symposium, Glasgow, June 20th, *Measuring mental health with machine learning and brain imaging*
- Programme national de recherche en IA, Rocquencourt, Jun 13th, *Infrastructures logicielles: Autour de scikit-learn*
- Rencontre Heka-Soda, PariSanté Campus, May 9th, *Motivation for the Soda team*
- Symposium on computational approaches to the mind, Oct 21st, online, *Model predictions advance science more than modeling ingredients*
- Keynote at Turing Workshop on Open-Source AI Software for Healthcare, Turing institute London, Nov 21st, *Open science for health data, Applying the scikit-learn recipe?*

Marine Le Morvan

- Service Now seminar, California, June 2022
  Learning with missing values: theoretical insights and application to health databases?
- IA and digital healthcare junior seminar, PariSante Campus, France, June 2022
  Learning with missing values: theoretical insights and application to health databases?
- Workshop on missing data and survival analysis, Angers, France, May 2022
  What's a good imputation to predict with missing values?
- Statistics seminar, University of Jyväskylä, Finland, April 2022
  What's a good imputation to predict with missing values?

- Rencontre Heka-Soda, PariSanté Campus, May 2022
  What's a good imputation to predict with missing values?

Jill-Jênn Vie

- Hong Kong ML meetup, Hong Kong, Jan 2022
  Machine learning for education & training
- Colloque sco-sup piloter le bac -3/+3, Institut des hautes études de l'éducation et de la formation (IH2EF), Poitiers, Feb 2022
  Vers des modèles prédictifs
- Soda seminar, Mar 2022
  Regression with sparse features and side information
- Rencontre Heka-Soda, PariSanté Campus, May 9
  Privacy-Preserving Tabular Data Generation
- Plateforme francophone IA (PFIA), Saint-Étienne, June 2022
  Fairness et confidentialité en IA pour l'éducation : risques et opportunités
- New in ML workshop, ICML 2022, Baltimore (remote), USA, June 2022
  Coding best practices
- Seminar, Kyoto University, Japan, Dec 12
  Modeling uncertainty for policy learning in education
- IEEE BigData 2022, Osaka, Japan, Dec 21
  Variational factorization machines for large-scale recommender systems

Tomas Rigaux

- European Conference on Technology-Enhanced Learning, EC-TEL, Sep 2022
  Privacy-Preserving Educational Data Generation

11.1.5 Leadership within the scientific community

Gaël Varoquaux Panel member on "table relational learning" NeurIPS 2022 workshops, New Orleans Dec 2nd.

Panel member on reproducibility in machine learning at ICLR, Apr 29th, remote

11.1.6 Scientific expertise

Gaël Varoquaux Member of the jury of the Udopia PhD grants.

Marine Le Morvan Member of the jury of the Udopia PhD grants.

11.1.7 Research administration

Jill-Jênn Vie

- Secrétaire de la Société informatique de France (SIF)
- Member of the CDT (commission de développement technologique) at Inria Saclay.

11.2 Teaching - Supervision - Juries

Courses

Gaël Varoquaux

- Machine learning for social sciences, EHESS, 8h
- Machine learning with missing values, ODSC, 3h

- Model Selection and validation in machine learning, Euroscipy (Basel), 1.5h
- Practicalities of machine learning: model selection and data preparation, Summer school Faaborg, 3h
- Model Selection for machine learning, Max Planck Academy, 1.5h

Marine Le Morvan

- Advanced Machine Learning, Ecole Polytechnique, 21h
- Deep Learning labs, Ecole Polytechnique, 28h

Olivier Grisel

- Deep Learning, Université Bretagne Sud, 21h

Judith Abecassis

- Causal inference, NYU Paris, 84h

Jill-Jênn Vie

- Préparation en IA à l'agrégation d'informatique, Sorbonne Université & ENS, 9h
- Deep learning: do it yourself, ENS, 37,5h (éq. TD)
- Advanced algorithms, École polytechnique, 14h

E-learning

Machine learning with Scikit-learn MOOC 40 hours of learning starting as an introduction to machine learning and covering more advanced topics such as data preparation and model selection. Accessible on inria.github.io/scikit-learn-mooc, and designed by Loïc Esteve, Arturo Amor, Guillaume Lemaître, Olivier Grisel, Gaël Varoquaux.

11.2.1 Supervision

Gaël Varoquaux Supervising the following PhD students: Samuel Brasil, Lihu Chen, Bénédicte Colnet, Alexis Cvetkov-Iliev, Matthieux Doutreligne, Alexandre Perez.

Also supervising undergraduate intern Anand-Arnaud Pajaniradjane and apprentice master Lilian Boulard.

Marine Le Morvan Supervising Alexandre Perez (PhD student).

Jill-Jênn Vie Supervising Samuel Brasil (PhD student), Tomas Rigaux (engineer)

11.2.2 Juries

Gaël Varoquaux Member of the following PhD juries: Léonard Blier (université Paris-Saclay, president), Jules Leguy (université Angers, president), Eric Daoud (université Paris Saclay, président), Balthazar Donon (université Paris Saclay, président) Tongxue Zhou (université de Caen, rapporteur).

Jill-Jênn Vie Member of the jury of the "agrégation d'informatique" and the entrance selective exam of ENS.

Marine Le Morvan Member of the following PhD juries: Florent Bascou (université de Montpellier, examinatrice).

11.3 Popularization

11.3.1 Internal or external Inria responsibilities

Gaël Varoquaux Member of the GPAI: Global Partnership on AI, a diplomatic mission to guide policy making around AI.

Jill-Jênn Vie Member of the European Commission Expert Group on AI & Data in Education & Training

11.3.2 Articles and contents

Gaël Varoquaux *"Beau parleur comme une IA"*, Fabien Suchanek & Gaël Varoquaux, The conversation, theconversation.com/beau-parleur-comme-une-ia-196084

Jill-Jênn Vie Éditeur du blog Binaire

11.3.3 Education

Gaël Varoquaux Course (1.5 hour, on model validation) at *AI for Ukraine* techukraine.org/2022/08/10/ai-for-ukraine-new-edtech-project-from-ai-house-to-support-the-ukrainian-tech-community/

11.3.4 Interventions

Gaël Varoquaux Presentation on AI to the executives of the interior minister, Inria Saclay, Sep 30th.

Loic Esteve Presentation on scikit-learn at the cyber campus, in the context of cyber security, Paris, Nov 21st.

Jill-Jênn Vie Café des sciences (university students; and elder customers from a library), Le Creusot, Apr 2022, *Algorithmes de recommandation : comment ça marche ?*

12 Scientific production

12.1 Major publications

[1] Y. Bergner, P. Halpin and J.-J. Vie. 'Multidimensional Item Response Theory in the Style of Collaborative Filtering'. In: *Psychometrika* 87.1 (Mar. 2022), pp. 266–288. DOI: 10.1007/s11336-021-09788-9. URL: https://hal.science/hal-03895623.

[2] L. Grinsztajn, E. Oyallon and G. Varoquaux. 'Why do tree-based models still outperform deep learning on typical tabular data?' In: NeurIPS 2022 Datasets and Benchmarks Track. Advances in Neural Information Processing. New Orleans, United States, 28th Nov. 2022. URL: https://hal.archives-ouvertes.fr/hal-03723551.

[3] G. Varoquaux and V. Cheplygina. 'Machine learning for medical imaging: methodological failures and recommendations for the future'. In: *npj Digital Medicine* 5.1 (Dec. 2022), p. 48. DOI: 10.1038/s41746-022-00592-y. URL: https://hal.archives-ouvertes.fr/hal-03818456.

[4] J.-J. Vie, T. Rigaux and H. Kashima. 'Variational Factorization Machines for Preference Elicitation in Large-Scale Recommender Systems'. In: IEEE BigData 2022. Osaka, Japan, 17th Dec. 2022. URL: https://hal.inria.fr/hal-03880215.

12.2 Publications of the year

International journals

[5] Y. Bergner, P. Halpin and J.-J. Vie. 'Multidimensional Item Response Theory in the Style of Collaborative Filtering'. In: *Psychometrika* 87.1 (Mar. 2022), pp. 266–288. DOI: 10.1007/s11336-021-09788-9. URL: https://hal.science/hal-03895623.

[6] D. Chyzhyk, G. Varoquaux, M. Milham and B. Thirion. 'How to remove or control confounds in predictive models, with applications to brain biomarkers'. In: *GigaScience* 11 (12th Mar. 2022). DOI: 10.1093/gigascience/giac014. URL: https://hal.inria.fr/hal-03607651.

[7] B. Colnet, J. Josse, G. Varoquaux and E. Scornet. 'Causal effect on a target population: a sensitivity analysis to handle missing covariates'. In: *Journal of Causal Inference* 10.1 (15th Sept. 2022), pp. 372–414. DOI: 10.1515/jci-2021-0059. URL: https://hal.science/hal-03473691.

[8] A. Cvetkov-Iliev, A. Allauzen and G. Varoquaux. 'Analytics on Non-Normalized Data Sources: more Learning, rather than more Cleaning'. In: *IEEE Access* 10 (2022), pp. 42420–42431. DOI: 10.1109/ACCESS.2022.3168013. URL: https://hal.archives-ouvertes.fr/hal-03647434.

[9] A. Cvetkov-Iliev, A. Allauzen and G. Varoquaux. 'Relational Data Embeddings for Feature Enrichment with Background Information'. In: *Machine Learning* (2022). URL: https://hal.archives-ouvertes.fr/hal-03848124.

[10] B. Hebling Vieira, F. Liem, K. Dadi, D. Engemann, A. Gramfort, P. Bellec, R. C. Craddock, J. Damoiseaux, C. Steele, T. Yarkoni, N. Langer, D. Margulies and G. Varoquaux. 'Predicting future cognitive decline from non-brain and multimodal brain imaging data in healthy and pathological aging'. In: *Neurobiology of Aging* 118 (Oct. 2022), pp. 55–65. DOI: 10.1016/j.neurobiolaging.2022.06.008. URL: https://hal.inria.fr/hal-03808175.

[11] A. Lamer, M. Fruchart, N. Paris, B. Popoff, A. Payen, T. Balcaen, W. Gacquer, G. Bouzillé, M. Cuggia, M. Doutreligne and E. Chazard. 'Standardized Description of the Feature Extraction Process to Transform Raw Data Into Meaningful Information for Enhancing Data Reuse: Consensus Study'. In: *JMIR Medical Informatics* 10.10 (2022), e38936. DOI: 10.2196/38936. URL: https://hal.archives-ouvertes.fr/hal-03818320.

[12] Y. Liu and G. Varoquaux. 'Understanding Brain Network Dynamics in Autism Begs for Generalization'. In: *Biological Psychiatry* 91.11 (June 2022), pp. 916–917. DOI: 10.1016/j.biopsych.2022.03.007. URL: https://hal.archives-ouvertes.fr/hal-03818457.

[13] R. Menuet, R. Meudec, J. Dockès, G. Varoquaux and B. Thirion. 'Comprehensive decoding mental processes from Web repositories of functional brain images'. In: *Scientific Reports* 12.7050 (Dec. 2022). DOI: 10.1038/s41598-022-10710-1. URL: https://hal.inria.fr/hal-03666470.

[14] P. Ortega-Ramírez, V. Pot, P. Laville, S. Schlüter, D. A. Amor-Quiroz, D. Hadjar, A. Mazurier, M. Lacoste, C. Caurel, V. Pouteau, C. Chenu, I. Basile-Doelsch, C. Henault and P. Garnier. 'Pore distances of particulate organic matter predict N 2 O emissions from intact soil at moist conditions'. In: *Geoderma* 429 (Jan. 2023), p. 116224. DOI: 10.1016/j.geoderma.2022.116224. URL: https://hal.inrae.fr/hal-03878855.

[15] I. Petria, S. Albuquerque, G. Varoquaux, J.-J. Vie, G. Velho, G. Perseghin, R. Roussel and L. Potier. '424-P: Body-Weight Variability and Risk of Cardiovascular Outcomes in Type 1 Diabetes: Results from the DCCT/EDIC Studies'. In: *Diabetes* 71.Supplement_1 (1st June 2022). DOI: 10.2337/db22-424-P. URL: https://hal.archives-ouvertes.fr/hal-03818459.

[16] N. Traut, K. Heuer, G. Lemaître, A. Beggiato, D. Germanaud, M. Elmaleh, A. Bethegnies, L. Bonnasse-Gahot, W. Cai, S. Chambon, F. Cliquet, A. Ghriss, N. Guigui, A. de Pierrefeu, M. Wang, V. Zantedeschi, A. Boucaud, J. v. D. Bossche, B. Kegl, R. Delorme, T. Bourgeron, R. Toro and G. Varoquaux. 'Insights from an autism imaging biomarker challenge: promises and threats to biomarker discovery'. In: *NeuroImage* 255 (2022), p. 119171. DOI: 10.1016/j.neuroimage.2022.119171. URL: https://hal.science/hal-03637273.

[17] G. Varoquaux and V. Cheplygina. 'Machine learning for medical imaging: methodological failures and recommendations for the future'. In: *npj Digital Medicine* 5.1 (Dec. 2022), p. 48. DOI: 10.1038/s41746-022-00592-y. URL: https://hal.archives-ouvertes.fr/hal-03818456.

International peer-reviewed conferences

[18] L. Chen, G. Varoquaux and F. Suchanek. 'Imputing out-of-vocabulary embeddings with LOVE makes language models robust with little cost'. In: ACL 2022 - 60th Annual Meeting of the Association for Computational Linguistics. Dublin, Ireland, 22nd May 2022. URL: https://hal.archives-ouvertes.fr/hal-03613101.

[19] L. Grinsztajn, E. Oyallon and G. Varoquaux. 'Why do tree-based models still outperform deep learning on typical tabular data?' In: NeurIPS 2022 Datasets and Benchmarks Track. Advances in Neural Information Processing. New Orleans, United States, 28th Nov. 2022. URL: https://hal.archives-ouvertes.fr/hal-03723551.

[20] S. Minn, J.-J. Vie, K. Takeuchi, H. Kashima and F. Zhu. 'Interpretable Knowledge Tracing: Simple and Efficient Student Modeling with Causal Relations'. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. 11. Vancouver, Canada, 15th Dec. 2021, pp. 12810–12818. DOI: 10.1609/aaai.v36i11.21560. URL: https://hal.inria.fr/hal-03895625.

[21] J.-J. Vie, T. Rigaux and H. Kashima. 'Variational Factorization Machines for Preference Elicitation in Large-Scale Recommender Systems'. In: IEEE BigData 2022. Osaka, Japan, 17th Dec. 2022. URL: https://hal.inria.fr/hal-03880215.

[22] J.-J. Vie, T. Rigaux and S. Minn. 'Privacy-Preserving Synthetic Educational Data Generation'. In: EC-TEL 2022 - 17th European Conference on Technology Enhanced Learning. Toulouse, France, 12th Sept. 2022. URL: https://hal.inria.fr/hal-03715416.

Conferences without proceedings

[23] A. Perez-Lebel, M. L. Morvan and G. Varoquaux. 'Beyond calibration: estimating the grouping loss of modern neural networks'. In: ICLR 2023 – The Eleventh International Conference on Learning Representations. Kigali, Rwanda, 2023. URL: https://hal.science/hal-03829870.

Scientific book chapters

[24] G. Varoquaux and O. Colliot. 'Evaluating machine learning models and their diagnostic value'. In: *Machine Learning for Brain Disorders*. 2022. URL: https://hal.archives-ouvertes.fr/hal-03682454.

Reports & preprints

[25] B. Colnet, J. Josse, G. Varoquaux and E. Scornet. *Reweighting the RCT for generalization: finite sample error and variable selection*. 4th Nov. 2022. URL: https://hal.archives-ouvertes.fr/hal-03822662.

[26] M. Doutreligne and G. Varoquaux. *How to select predictive models for causal inference?* 19th Jan. 2023. URL: https://hal.science/hal-03946902.